2024

AI Text-To-Image Tools: Evaluating Risks of Defamation

Olivia Mao

AI Text-To-Image Tools: Evaluating Risks of Defamation

I. Introduction

Imagine receiving your wedding photos only to discover an unwanted person in the background. Consider being tasked by your supervisor at the last minute to create a new logo for a new company marketing campaign. Envision the need to capture a vertical picture and expand the background to fit a horizontal poster. Faced with similar challenges twenty years ago, one would likely have to seek the expertise of a professional designer or navigate complicated image editing tools. However, picture a scenario in which access to an AI tool could mitigate these challenges. AI Text-to-image tools, such as Photoshop's Generative Fill or Open AI's DALL-E, can resolve these issues with a few clicks. How might these technologies change photo editing, graphic design, and image assessment?

The innovative AI Text-to-image technology, led by tools like Photoshop's Generative Fill, provides a powerful means to effortlessly expand or modify original photos. AI algorithms can learn to expand or alter the original image, either through user-initiated commands or through its own analysis.[1] However, the widespread adoption of this technology raises significant legal concerns. Text-to-image tools have the potential to inflict harm upon individuals, including issues such as libel, slander, and defamation. Creators can exploit Text-to-image tools to generate deceptive images to communicate false and offensive messages about the targeted individual. However, the existing defamation framework is insufficient to protect victims of AI Text-to-image tools, and there is a need for legislation to adjust the intent requirement. Specifically, defamation facilitated by features resembling deepfakes should be regulated as a subcategory of deepfake; defamation though AI-generated content through text instructions,

[1] Adobe, *Tap Into The Power Of AI Photo Editing*, https://www.adobe.com/products/photoshop/ai.html.

incorporating both offensive and false elements, should create a presumption that the intent requirement for defamation is fulfilled. This presumption will enhance the protection of victims against defamation through Text-to-image tools.

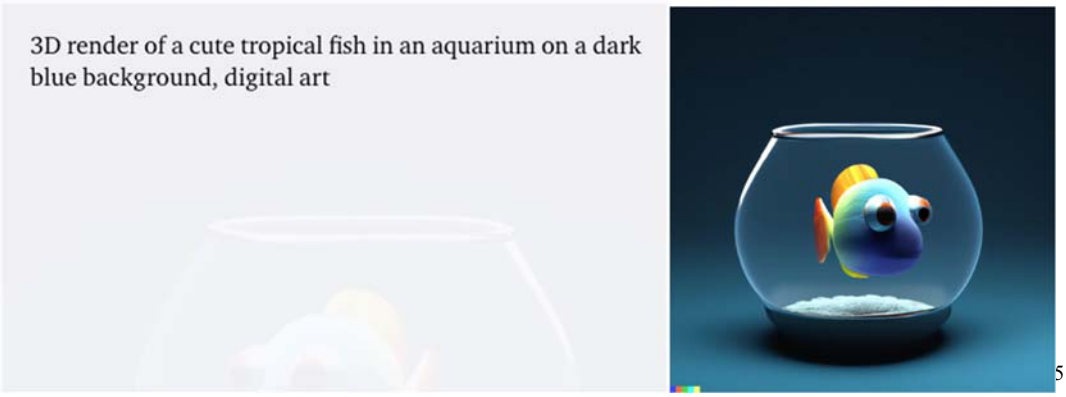
Part II provides an overview of the Text-to-image technology and explains the background surrounding this powerful AI technology. Part III points out that the issue arising from Text-to-image technology lies in its ability to facilitate defamation, rendering image editing more accessible, cost-effective, and realistic. Part IV introduces and conducts a legal analysis of the elements of defamation. Part V argues that the current regulatory framework governing Text-to-image tools is inadequate and advocates for changes in defamation's intent requirement for AI Text-to-image tools.

II. AI Text-To-Image Background

AI Text-to-image models involve users providing a text description into a textbox, and these models will generate images that match the description as closely as possible.[2] Descriptions can range from simple ones such as "an apple" or "a cat sitting on a couch" to more complex ones such as "a cute sloth holding a small treasure chest. A bright golden glow is coming from the chest."[3] Tools such as OpenAI's Dall-E 2, Stability AI's Stable Diffusion, and Midjourney Lab's Midjourney image generator create the opportunity to generate synthetic imagery solely based on text prompts, without the need for feeding the system preexisting pictures and videos.[4] Undoubtedly, tools such as DALL-E and Midjourney represent a revolutionary step in creating

[2] Yonghui Wu, David Fleet, *How AI Creates Photorealistic Images From Text* (June 22, 2022), https://blog.google/technology/research/how-ai-creates-photorealistic-images-from-text/.
[3] *Id*.
[4] Jon M. Garon, *An Ai's Picture Paints A Thousand Lies: Designating Responsibility For Visual Libel*, https://www.journaloffreespeechlaw.org/garon.pdf.

realistic and creative images rapidly and straightforwardly. They present an ideal solution for everyday graphic design and image creation needs.



[5]

OpenAI's Dall-E's image generation (right) based on the text prompt (left)

AI Text-to-image tools become more powerful when they are combined with image editing, as seen in tools such as Photoshop's Generative Fill. Generative Fill, powered by Adobe's Generative AI model "Adobe Firefly", allows creators to incorporate AI-generated image into the original image. Even though Adobe Photoshop may not have always been the most intuitive image-editing tool for average users, Generative Fill is designed to expand creator's creative expression and productivity by using natural language and concepts to generate digital content within seconds.[6] Through its power to automatically match the perspective, lighting, and style of images, Generative Fill allows users without years of experience to achieve remarkable results.[7]

Adobe's Generative AI model, Adobe Firefly, employs the diffusion model.[8] The new diffusion model is based on "the use of adding noise to the images to train the system to identify

[5] DALL-E, https://labs.openai.com/.
[6] Adobe, *Adobe Unveils Future of Creative Cloud With Generative AI as a Creative Co-Pilot in Photoshop*, https://news.adobe.com/news/news-details/2023/Adobe-Unveils-Future-of-Creative-Cloud-with-Generative-AI-as-a-Creative-Co-Pilot-in-Photoshop-default.aspx/.
[7] *Id*.
[8] Adobe, *Adobe Firefly vs. Stable Diffusion: Bring more ideas into your workflows quickly with Firefly*. https://www.adobe.com/sensei/generative-ai/discover/firefly-vs-stable-diffusion.html.

visual elements from the competing data."[9] In simple terms, the diffusion model works by imitating the data they have been trained on.[10] The diffusion model is similar to a master chef recreating a dish based on one that he or she has tasted before: the chef tastes a dish, understands the ingredients, and subsequently prepares a similar new dish.[11] Thus, diffusion models can learn to generate new images similar to the ones they have been trained on.[12]

Photoshop's Generative Fill can edit images through four main features: generating new items, changing backgrounds, expanding images, and removing unwanted elements. When generating new items, users select an area within the image and input text instructions. Generative Fill will generate the item according to the text instructions within seconds.[13] Generative Fill to create new items that blend with the original image through simple instructions such as "a lighthouse", "road", "Jaguar drinking from pond". For changing the background, users can select the background and enter instructions in the textbox to transform it from a local city street to a location across the globe.[14] Expanding an image beyond its borders is straightforward as well. By dragging the Crop tool beyond the image's original border to the desired size, Generative Fill will add contents that fit the original image.[15] Finally, if users need to remove unwanted elements, they can leave the textbox blank, and Generative Fill can remove those elements. Generative Fill can swap out obstructive elements such as telephone lines with unobstructed sky, or it can replace stray strangers with flowers and foliage.[16]

---

[9] Altexsoft, *AI Image Generation Explained: Techniques, Applications, and Limitations* (Jul. 9, 2023), https://www.altexsoft.com/blog/ai-image-generation/#:~:text=Stable%20Diffusion%20utilizes%20the%20Latent,with%20the%20textual%20description%20provided.
[10] *Id*.
[11] *Id*.
[12] *Id*.
[13] Adobe, *Tap Into The Power Of AI Photo Editing*, https://www.adobe.com/products/photoshop/ai.html.
[14] *Id*.
[15] *Id*.
[16] *Id*.

[17]

Generative Fill adding a lighthouse that blends with the original image

Generative Fill, contrary to past perceptions of Photoshop, is extremely user-friendly. Users start by selecting an object or area within the image and then clicking the Generative Fill button in the Task Bar. Depending on their objective, users can either input a text prompt or leave the prompt box blank. After clicking "Generate", thumbnail previews of variations generated based on the prompt will become available. The Generative Layer will be created as a separate non-destructive layer independent of the original image.[18] Users can click "Generate" repeatedly until they reach the desired result.

Given these functions, Generative Fill has the potential to be powerful and revolutionary across various industries, including art design, advertising, media, and marketing. Generative Fill's applications can be diverse as well. For example, in the realm of the art design, it can be used to expand photographs taken during special occasions, incorporate materials such as "tiger fur" onto texts to create unique fonts, and add color themes such as "80s neon" to existing images.[19] In marketing, it can transform product packaging designs, create logos for companies,

[17] Adobe, *Tap Into The Power Of AI Photo Editing*, https://www.adobe.com/products/photoshop/ai.html.
[18] Adobe, *Photoshop features Generative Fill. Now in Photoshop*. https://www.adobe.com/products/photoshop/generative-fill.html.
[19] Adobe, *Adobe Firefly: Out of Beta & Ready For Your Imagination*, YouTube (Sept. 13, 2023), https://www.youtube.com/watch?v=NPJNPrshhTo.

and design event posters.[20] Generative Fill opens up possibilities for integrating any creative idea to a part of an image.

III. Harms Of AI Text-To-Image Tools

Generative Fill introduces a new approach to image editing that simplifies the process in unprecedented ways. However, similar to many other groundbreaking AI tools, Generative Fill raises concerns related to discrimination, misuse, and misinformation.[21] This paper focuses specifically on misinformation that causes potential harm to individuals, including issues such as libel, slander, and defamation.

A. Harms Caused By AI Text-To-Image Tools

According to the Restatement of Torts, a communication is defamatory if "it tends so to harm the reputation of another as to lower him in the estimation of the community or to deter third persons from associating or dealing with him."[22] Therefore, creators can use various methods through AI Text-to-image tools to defame others, in order to harm the reputation of the victim and dissuade others from associating with the victim. These defamatory messages can be achieved by face-swapping the victim to an existing image, showing the victim with offensive items, or suggesting that the victim is in an inappropriate location. All these methods involve generating new images that communicate certain false messages about the victim.

First, Generative Fill can replace the individual's face in an image with the victim's face and adjust the overlapping areas to create a realistic representation. This adjustment can create the illusion that the victim is the individual engaged in inappropriate behavior in the image. For

[20] Adobe, *Adobe Firefly: Out of Beta & Ready For Your Imagination*, YouTube (Sept. 13, 2023), https://www.youtube.com/watch?v=NPJNPrshhTo.
[21] Charlotte Bird, Eddit L. Ungless, Atoosa Kasirzadeh, *Typology of Risks of Generative Text-to-Image Models*, AIES '23: Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society, Aug. 2023, at 396-410.
[22] Restat 2d of Torts, § 559.

example, in a tutorial using Generative Fill to create an image with Taylor Swift, the creator shows how smoothly Generative Fill can be used to replace the original person in the image. After replacing the person next to Taylor Swift with the creator in the original image, Generative Fill can then refine various details, such as the interaction between the individual's clothing, lighting, and shadows.[23] These adjustments significantly enhance the overall quality and completeness of the image.

While Generative Fill can be used creatively to generate images with entertainment values through face-swapping, it can also be used to generate harmful content such as revenge pornography. Creators can swap the victim's face onto existing pornography images and distribute those manipulated images with the intent to harm the victim. The reputational harm through face-swapping can extend to swapping the victim's face onto other types of existing offensive images. This act can harm the victim by convincing the community that the victim was, indeed, the person depicted in the original image.

Also, Generative Fill can be used to defame the victim by depicting the victim being in a specific location. This tool can expand the background of the original image, either with or without text prompt instructions. Through this feature, Generative Fill can easily transform an upper-body picture into a full-body picture. It can also extend a headshot to include hand gestures and reveal the surrounding context. For example, an upper-body picture can be expanded to include elements like hospital signs and patient chairs in the background. The manipulated image can communicate the false message that a victim is in the hospital and undergoing a diagnosis of an infectious disease.

[23] PiXimperfect, *Photoshop Generative Fill - 20 EPIC Uses, SUPERFAST!*, YouTube (May 30, 2023), https://www.youtube.com/watch?v=NvUZIm083P8.

Alternatively, a creator can use Generative Fill to communicate a false message about a person's activity by selecting any part of the image and providing text instructions. The creator can select an area within the image to either add new objects or substitute the area with different objects. AI-generated images can become so realistic that they are indistinguishable from real photos.[24] Therefore, any added object or incorporated street view can be realistic enough to convince others that it exists in reality. For instance, an image of an individual with relaxed hands can be manipulated to depict an individual with clenched fists or arms in a fighting position, suggesting that the individual is engaging in violent behaviors.

Regardless of the specific method, all these techniques have the potential to harm the victims' reputation across various aspects of their daily lives. The altered images could depict the victims violating religious taboos, harming their spouses or children, engaging in violence, conspiring with competitor companies, having a confections disease, engaging in cheating, participating in fabricated sexual activities, committing crimes, and more.[25] These images, once started disseminating within the community, could lead to the victims losing their jobs or opportunities, creating distrust between families, affecting their relationships friends and family, or even imposing a sense shame in themselves. These examples of defamatory images can be easily achieved by entering text instructions yet could be devastating and destructive to average individuals in the course of their normal lives.

B. How AI Text-To-Image Tools Facilitate Defamation

---

[24] Gary Grossman, *AI text-to-image processors: Threat to creatives or new tool in the toolbox?*, VentureBeat (Sept. 10, 2022, 6:40 AM), https://venturebeat.com/ai/ai-text-to-image-processors-threat-to-creatives-or-new-tool-in-the-toolbox/.

[25] *See* Christy Bieber, J.D., Adam Ramirez, J.D., What Is Defamation Of Character? Definition & Examples, ForbesAdvisor (May 19, 2023, 10:31 AM), https://www.forbes.com/advisor/legal/personal-injury/defamation-of-character/#:~:text=But%2C%20under%20a%20legal%20document,with%20their%20business%20or%20profession.

Manipulating new images is not a new concept and has existed before the creation of tools such as Photoshop's Generative Fill. However, the issues created by Generative Fill are significant because AI Text-to-image tools can facilitate defamation by allowing an ordinary computer user to manipulate images at a professional level within a few clicks. As AI Text-to-image technology advances rapidly, it facilitates libel, slander, and defamation in three different ways: simplified processes, creative content, and realistic images.

First, Text-to-image tools make creating false and defamatory images easier. Text-to-image model allows creators to generate original images from text-based prompts rather than requiring creators to input a source image.[26] Before Text-to-image tools, creators had to locate preexisting images of the item they wished to add to the new image. For example, when using traditional Photoshop to add an apple to an image, creators typically had to locate a pre-existing image featuring an apple that they liked. They would then select, crop, and drag the apple from the pre-existing image to the new image. Subsequently, creators need to smooth out the edges where the apple touches the background manually. The final outcome depends heavily on the creator's skill and experience.

Now, Generative Fill provides the opportunity to effortlessly replace the entire background or specific items within the image by selecting the area and entering corresponding text instructions. It can also adjust unnatural areas by selecting the area and clicking on the "Generate" button, without the need to individually edit each unnatural area. These functions significantly reduce the cost and complexity of engaging in defamation. Creators do not need to become proficient in Photoshop's complicated features to crop objects, drag items, or replace a

[26] Jon M. Garon, *An Ai's Picture Paints A Thousand Lies: Designating Responsibility For Visual Libel*, https://www.journaloffreespeechlaw.org/garon.pdf.

person from one image to another. With Generative Fill, creators can achieve the same results by typing in instructions, making knowledge of advanced photo editing skills unnecessary.

Second, defamation can manifest in unanticipated ways because images can now be more creatively manipulated. Unlike the traditional image-generating method that relies on pre-existing material, the current approach is based on text instructions. The more creative the text prompt that creators type in, the more creative the generated content will be. The instruction could encompass virtually anything that can be expressed in words, without worrying about whether the text prompt is realistic or logical. For example, it is possible to request putting a jaguar and a pond in a library, a scenario that is not realistic or feasible in real life.[27]

Furthermore, while expanding the image, creators can put in instructions or leave the text prompt box blank, allowing the AI algorithm to decide the content. If creators leave the text prompt box blank, the generated image is shaped by AI's training input, which may lead to the creation of unexpected images that humans may not have been able to image. Creators can generate new sets of content repeatedly until they are satisfied with the results and achieve the desired level of creativity. Therefore, creators can either use the power of words or let AI generate content to defame others in more unexpected ways.

Third, the Text-to-image generated images show a higher degree of realism, which makes defamation more compelling. Before AI Text-to-image tools, images were typically edited individually. This step requires the creators to pay attention to detail when adjusting the imperfections. The extent to which details could be adjusted depends on the experiences of the image creators. Creators often faced challenges in addressing every detail during the image editing process.

---

[27] Adobe, *Photoshop features Generative Fill. Now in Photoshop*. https://www.adobe.com/products/photoshop/generative-fill.html.

Now, Generative Fill can manage the minor details that humans tend to overlook. For example, in Photoshop's video "*Introduction to AI Generative Fill*", the creator demonstrates the powerful and logical decisions made by Generative Fill when generating content.[28] Generative Fill can calculate the position of the sun when extending the background of a mountain view image. When generating a reflection pool, Generative Fill did not forget to generate the bottom of the car within the reflection pool.[29] Creators, especially those with limited experience, tend to neglect these details, leading to a reduction in the overall completeness of the image. Furthermore, Generative Fill can be used to refine edges when swapping faces or items so that the image can pass as a real photograph to the average human eye. For example, when replacing a person in a family portrait, the area where two people's hair touches can look unnatural because of the cropping.[30] Creators can select this problematic area and use Generative Fill to generate a new area that seamlessly blends these two people's hair.[31] Creators can keep generating until they find a result that best fits the overall image.[32]

Even though companies like Adobe and Google are taking steps to implement watermarks on images to differentiate between real pictures and AI-generated pictures, it remains extremely challenging for the public to distinguish between the two.[33] As a result,

---

[28] Adobe Photoshop, *Introduction to Generative Fill | Adobe Photoshop*, YouTube (May 23, 2023), https://www.youtube.com/watch?v=Sp6K3qpVFO0.
[29] *Id*.
[30] PHLEARN, *How to Swap Faces in Photoshop Using AI Generative Fill*, YouTube (Sept. 29, 2023), https://www.youtube.com/watch?v=9YFDOoM7I5Q.
[31] *Id*.
[32] *Id*.
[33] Shirin Ghaffary, *What will stop AI from flooding the internet with fake images?* Vox ((June 3, 2023, 7:00 AM), https://www.vox.com/technology/23746060/ai-generative-fake-images-photoshop-google-microsoft-adobe

defaming others becomes easier using tools such as Generative Fill because the public usually perceives the content in the manipulated image as real items.[34]

IV. Defamation Claims By Victims Of AI Text-To-Image Tools

A. Legal Standard of Defamation

The First Amendment of the U.S. Constitution states that "Congress shall make no law. . . abridging the freedom of speech."[35] Defamation exists as an exception to the broad protection granted by the First Amendment.[36] Defamation laws and First Amendment protections can sometimes be in tension because there is an interest in promoting artistic creation while protecting individuals from harmful speech. In *Chaplinsky v. New Hampshire*, the supreme court recognized that "it is well understood that the right of free speech is not absolute at all times and under all circumstances. . . . These include the lewd and obscene, the profane, the libelous, and the insulting or 'fighting' words -- those which by their very utterance inflict injury or tend to incite an immediate breach of the peace."[37]

To create liability for defamation, the elements are: "(a) a false and defamatory statement concerning another; (b) an unprivileged publication to a third party; (c) fault amounting at least to negligence on the part of the publisher; and (d) either actionability of the statement irrespective of special harm or the existence of special harm caused by the publication."[38] Publication of defamatory matter occurs when there is communication intentionally or by a negligent act to one other than the person defamed.[39] Privileged publications include true

---

[34] *See* Gary Grossman, *AI text-to-image processors: Threat to creatives or new tool in the toolbox?*, VentureBeat (Sept. 10, 2022, 6:40 AM), https://venturebeat.com/ai/ai-text-to-image-processors-threat-to-creatives-or-new-tool-in-the-toolbox/.
[35] U.S. Const. amend. I.
[36] *Chaplinsky v. New Hampshire*, 315 U.S. 568, 571-72, 62 S. Ct. 766, 769 (1942).
[37] *Id.*
[38] Restat 2d of Torts, § 558.
[39] *Id*. at § 577.

statements, consent of another to the publication of defamatory matter concerning him, or other absolute privileges irrespective of consent (including judicial officers, attorneys at law, and parties to judicial proceedings).[40] Whether the defamation would be per se actionable may depend on the state. For example, a per se defamatory can be "[w]ords imputing a lack of chastity to any female or male".[41] Some other defamation actionable per se according to the Restatement includes criminal offense, loathsome disease, and matter incompatible with his business, trade, profession, or office.[42] In simpler words, Defamation requires the publication of a false fact that harms the reputation of another.[43] Both libel and slander fall under the broader category of defamation.[44]

There is already a pending case against OpenAI via the text-based Generative AI tool, ChatGPT.[45] In June 2023, radio host Mark Walters filed a libel case against OpenAI after ChatGPT generated a fabricated complaint containing allegations against Walters for fraud and embezzlement. ChatGPT fabricated the case which Walters was the defendant in response to a command to provide a summary of a pending civil rights lawsuits.[46] Defamation lawsuits involving AI tools, including Text-to-image tools, are likely to become increasingly prevalent in the future as these tools gain increased usage.

Besides text-based defamation, images can be defamatory as well. Older case laws recognize defamation based on manipulated images. For example, in *Kiesau v. Bantz,* an Iowa

---

[40] Restat 2d of Torts, § 581A, 583, 585-92A; See *Milkovich v. Lorain Journal Co.*, 497 U.S. 1, 13, 2703 (1990) (the privilege of "fair comment" is an affirmative defense to defamation in common law).
[41] Mich. Comp. Laws § 600.2911(1); *See Ogle v. Hocker*, 430 F. App'x 373, 374 (6th Cir. 2011).
[42] Restat 2d of Torts, § 570.
[43] Matthew B. Kugler, Carly Pace, Article, *Deepfake Privacy: Attitudes And Regulation*, 116 Nw. U.L. Rev. 611 (2021).
[44] Rodney A. Smolla, *Law of Defamation* § 1:10 (2d ed. 2008).
[45] *Recent Trends In Generative Artificial Intelligence Litigation In The United States*, https://www.klgates.com/Recent-Trends-in-Generative-Artificial-Intelligence-Litigation-in-the-United-States-9-5-2023
[46] *Walters v. OpenAI*, Case No. 23-cv-03122 (N.D. Ga. 2023).

court held that an altered image depicting a female police officer with her breasts exposed was libel per se.[47] In *Morsett v. "The Final Call"*, the New York Appellant Court upheld the jury verdict of libel from an image making a woman appear to be a convict in the newspaper.[48]

Defamation laws establish different levels of protection for public figures and private individuals. The distinctions are based on the different standards that must be met to establish liability. The landmark case *New York Times Co. v. Sullivan* set the standard for proving defamation for public figures. The *Sullivan* case involved a piece of writing by the New York Times against Sullivan, the commissioner of Montgomery, Alabama.[49] The court in *Sullivan* introduced the "actual malice test" for public figures, and stated that the standard "prohibits a public official from recovering damages for a defamatory falsehood relating to his official conduct unless he proves that the statement was made with 'actual malice'—that is, with knowledge that it was false or with reckless disregard of whether it was false or not."[50] The court thus held that public figures need to prove "actual malice" to succeed in a defamation suit.[51] To prove actual malice, a plaintiff must "demonstrate with clear and convincing evidence that the defendant realized that his statement was false or that he subjectively entertained serious doubts as to the truth of his statement."[52]

Subsequent case laws have extended greater protection to private individuals who had not voluntarily placed themselves in the public eye.[53] States may define the standard of liability for

---

[47] *Kiesau v. Bantz,* 686 N.W.2d 164, 178 (Iowa 2004).
[48] *Morsette v. "The Final Call,"* 764 N.Y.S.2d 416 (N.Y. App. Div. 2003).
[49] *N.Y. Times Co. v. Sullivan*, 376 U.S. 254, 280 (1964).
[50] *Id.*
[51] *Id*.
[52] *Bose Corp. v. Consumers Union of U.S., Inc.*, 466 U.S. 485, 511 (1984), *Khawar v. Globe Internat.*, 965 P.2d 696, 709 (1998).
[53] *Gertz v. Robert Welch*, 418 U.S. 323, 352 (1974) (state law could properly impose liability on a less demanding showing on a lawyer who was neither a public official nor a public figure); *Time, Inc. v. Firestone*, 424 U.S. 448, 453, 96 S. Ct. 958, 965 (1976) (court found that respondent was not a public figure)

defamation against private individuals.[54] Various state laws indicate that while public figures require an actual malice standard, private individuals only need to demonstrate that the defendant acted negligently, which is a lower standard than actual malice. [55] To prove the negligence standard of fault requires proof of failure to exercise reasonable care.[56] The conduct of the defendant is to be measured against what a reasonably prudent person would have done under the same or similar circumstances.[57] However, to recover punitive damages, in addition to actual damages, some states still require the actual malice standard even for actions brought by private individuals.[58]

However, a defense of defamation includes parody, which may serve as an absolute defense in defamation suits. For something to be considered a parody, it must be perceived as a parody by its audience.[59] In *Hustler Magazine v. Falwell*, the court dismissed the defamation claim because the content was ruled as parody.[60] The court in *Hustler Magazine* decided that the content must not reasonably be understood as describing actual facts… or events." [61] Therefore, if content generated by AI Text-to-image tools is established as a parody, it can be an effective defense against defamation claims.

B. Defamation Claims Against Both The Creator And The Designer

---

[54] *Brown v. Kelly Broad. Co.*, 48 Cal. 3d 711, 722-23 (1989).
[55] *Krass v. Obstacle Racing Media, LLC*, No. 1:19-CV-5785-JPB, 2023 U.S. Dist. LEXIS 47073, at *57 (N.D. Ga. Mar. 21, 2023) (plaintiff who is a private figure "must prove that the defendant acted with ordinary negligence. A plaintiff who is a public figure, on the other hand, must meet a more stringent standard of liability"); *Brawley v. Rouhfar*, No. 65399-7-I, 2011 Wash. App. LEXIS 1718, at *21 (Ct. App. July 25, 2011) (a "private figure defamation plaintiff need prove only the defendant's negligence by a preponderance of the evidence to establish a prima facie case of defamation"); *Khawar v. Globe Internat.*, 19 Cal. 4th 254, 274, (1998) ("In California, this court has adopted a negligence standard for private figure plaintiffs seeking compensatory damages in defamation actions")
[56] *Guntheroth v. Rodaway*, 727 P.2d 982, 986 (1986).
[57] *Kostenko v. CBS Evening News*, 265 F. Supp. 3d 672, 680 (S.D. W. Va. 2017).
[58] *Khawar v. Globe Internat.*, 19 Cal. 4th 254, 274, (1998)
[59] 50 Am. Jur 2d, Libel and Slander 156 (2018).
[60] *Hustler Magazine v. Falwell*, 485 U.S. 46, 57 (1988).
[61] *Id*.

Under the existing framework, a claim of defamation against the creator of the edited images may offer the plaintiff the chance to recover some damages. However, such a claim is generally unsuccessful because it cannot provide adequate compensation for the plaintiff. The creators usually have few resources and are hard to locate. However, a claim against the designer of the AI Text-to-image tool is often unsuccessful because it is challenging for the plaintiff to meet the intent requirement to establish liability for the designer.

To establish a defamation claim against the creator or the designer, the plaintiff needs to demonstrate that (a) the manipulated image expresses a false statement (b) the image was unprivileged and published (c) there is the requisite degree of fault regarding the truth of the statement (negligence standard for private individuals or actual malice standard for public officials) (d) the victim experienced harm and damages, either per se harm or special harm caused by the publication.[62]

A typical defamation scenario caused by AI Text-to-image tool generally involves a victim encountering a posting of a false image about the victim. The image could involve editing of the victim's pre-existing picture or face-swapping the victim to another image. It is usually straightforward to prove that the image creates a false statement because the image was edited to include incorrect persons, fake items, or fabricated backgrounds. The plaintiff can show the falsity of the image by showing that the image was edited to include elements that were not in the original image. Also, proving that the image is published to someone other than the person defamed is usually a low threshold. A typical victim would see the circulation of the defamatory image either online or in the community before the victim brings action. Third, the plaintiff needs to show the harm suffered, which can be proved by the impact on the reputation within the

[62] *See Ogle v. Hocker*, 430 F. App'x 373, 374 (6th Cir. 2011) (element of defamation in Michigan); *Bedford v. Spassoff*, 520 S.W.3d 901, 904 (Tex. 2017) (element of defamation in Texas).

community. Fourth, the plaintiff needs to prove the actual malice standard if the plaintiff is a public figure or the negligence standard if the plaintiff is a private individual. Because the creator knowingly altered the original image to create the new image, the creator must have known or at least recklessly disregarded the creation of false content when purposefully using Text-to-image tools to manipulate an image. Since the negligence standard is lower than the actual malice standard, if the actual malice standard is fulfilled, the negligence standard can be fulfilled as well. Therefore, for actions against creators, the victim may be able to recover some damages from the creator of the defamatory images.

However, pursuing the claim against the creator of the content is not always practical. Early defamation cases usually involved the publishing of content in newspapers or magazines, and the editor or author was typically easy to locate.[63] In the digital era, defamation cases can involve posting with an online account or social media account.[64] While it is certainly possible to locate the publisher or creator if the owner of the account can be identified, it is not always guaranteed. On the other hand, the designer of the AI Text-to-image tools, especially big corporations like Adobe, will have more resources than individuals. As a result, the plaintiff is more likely to recover full damages if they pursue a claim after the designers. Therefore, bringing an action against the corporation that created the AI Text-to-image tool may be the better avenue for the plaintiff.

The plaintiff will encounter significant challenges when bringing a defamation claim against the designer or the developer of the AI tool. The biggest obstacle will be the intent

[63] *Schafer v. Time, Inc.*, 142 F.3d 1361, 1364 (11th Cir. 1998) (defendant is a magazine publisher); *Paterson v. Little, Brown & Co.*, 502 F. Supp. 2d 1124, 1127 (W.D. Wash. 2007) (defendant is a publisher of series of essays); *N.Y. Times Co. v. Sullivan*, 376 U.S. 254, 280 (1964). (defendant is a New York corporation which publishes the New York Times, a daily newspaper)

[64] *See Bedford v. Spassoff*, 520 S.W.3d 901, 904 (Tex. 2017) (defamation involving a Facebook post).

requirement. Even if the plaintiff can prove the other elements of defamation, including the falsity of the image, unprivileged and publication, harm, and damages, it is difficult for plaintiffs to prove the requisite intent. Under the actual malice standard, the designer of the tool needs to have knowledge that the content was false or recklessly disregard whether the content was false. The designer likely does not have knowledge of the falsehood and companies such as Adobe are likely unable to identify falsehoods in every creator's image. Thus, the plaintiff cannot prove the actual malice standard absent extreme circumstances. Even under the negligence standard, the plaintiff will face challenges in establishing that the designer fails to exercise reasonable care in their design. For example, Adobe's "Generative AI User Guideline" limits the creation of illegal content or content that violates the rights of others and reviews through automated and manual methods of prompts and the results generated by generative AI features for abuse prevention and content filtering purposes.[65] These procedures demonstrate that the designer is exercising some degree of care in maintaining the tool. The burden on the plaintiff to prove a reasonable degree of care is challenging because the plaintiff usually lacks access to the inside information of how the designer company is filtering and preventing harmful content. Therefore, without fulfilling the intent requirement, the plaintiff cannot recover damages from the designer. This limitation will probably deny the plaintiff a significant portion of the relief sought in the defamation case.

V. Proposed Regulations

Despite AI Text-to-image tools' ability to assist in everyday tasks, their risks cannot be overlooked. There is an increasing need for regulations to mitigate potential harms. While facing defamation risks from Text-to-image tools, victims should be able to exercise their rights to

[65] Adobe, *Adobe Generative AI User Guidelines*, https://www.adobe.com/legal/licenses-terms/adobe-gen-ai-user-guidelines.html#:~:text=Do%20not%20use%20Adobe's%20generative,Pornographic%20material%20or%20explicit%20nudity

protect their reputations from being harmed as much as possible. The legislature should consider methods for safeguarding the public from defamation as technology continues to evolve.

The proposed regulations are based on different features of Text-to-image tools. AI Text-to-image tools have features that resemble deepfakes and features that are beyond those of deepfakes. These two different types of features warrant different standards. For work made by Text-to-image tools in a way that resembles deepfakes, Text-to-image tools should be subject to regulations within the established framework of deepfakes. For those work created by features beyond those of deepfakes, AI Text-to-image tools need stricter regulations than deepfakes because Text-to-image tools are more likely to be abused.

A. Text-To-Image Features That Are Similar To Those Of Deepfakes

Before exploring the overlapping features between deepfakes and AI Text-to-image tools, it is essential to provide a brief background on what deepfakes are. Deepfakes are "manipulated videos, or other digital representations produced by sophisticated artificial intelligence, that yield fabricated images and sounds that appear to be real."[66] Deepfake can be extremely complex, but in its most basic form, creators need a "base video" and several source images of the face of the individual being replaced in the video.[67] Creators only need to click a few buttons to create the deepfake video.[68] What was once a complicated and expensive process has now become accessible to average creators at their fingertips, thanks to the help of AI algorithms.[69]

Deepfake technology has been used broadly in the movie industry. For instance, in the famous movie *Forrest Gump,* the movie used deepfake technology to enable the main character,

[66] Lindsey Wilkerson, Note, *Still Waters Run Deep(fakes): The Rising Concerns of "Deepfake" Technology and Its Influence on Democracy and the First Amendment*, 86 Mo. L. Rev. 407, 408 (2021).
[67] Erik Gerstner, Face/Off: "DeepFake" Face Swaps and Privacy Laws, https://www.iadclaw.org/assets/1/17/Face_Off_-_DeepFake_Face_Swaps_and_Privacy_Laws.pdf?4179 (2020).
[68] *Id*.
[69] *Id*.

Forrest Gump, to interact with President Kennedy.[70] However, deepfakes also subject individuals to defamatory risks and significant damage to personal reputation. For example, celebrities, presidents, and politicians have often been the target of deepfakes due to their media worthiness.[71] Other rising issues of deepfakes include bullying, revenge porn, video manipulation, audio manipulation, and extortion.[72] Celebrities such as Gal Gadot, Masie Williams, and Daisy Ridley have all been victims of deepfake pornography, in which the creator superimposed their pictures onto the bodies of adult video stars in pornographic films.[73]

The current regulation governing defamation through the use of deepfakes is complicated. The creation of digital content, such as deepfakes, raises First Amendment issues. Creators of deepfakes may argue that their work is a form of self-expression.[74] They may also argue that any restriction on the use of deepfakes would infringe upon their freedom of speech and expression, constituting a First Amendment violation.[75] However, defamation lawsuits can be used as a means of seeking redress for deepfake victims. In theory, when an individual becomes a target of a deepfake video that might harm his or her reputation, the individual can bring a defamation claim against the creator of the deepfake video. Generally speaking, videos and images and treated the same under the law.[76] Thus, defamation laws that recognize defamation through images also apply to defamation through videos. The harm of defamation

---

[70] Forrest Gump (Paramount Pictures 1994).
[71] *See* The Times, *Deepfakes of Donald Trump 'arrest' spread on social media*, https://www.thetimes.co.uk/article/donald-trump-deepfakes-ai-twitter-g50n7vnbm
[72] Lourdes Vazques, *Recommendations for regulation of deepfakes in the U.S.: Deepfake laws should protect everyone not only public figures*, https://www.ebglaw.com/assets/htmldocuments/uploads/2021/08/Reif-Fellowship-2021-Essay-2-Recommendation-for-Deepfake-Law.pdf
[73] Russell Spivak, *Article: "Deepfakes": The Newest Way To Commit One Of The Oldest Crimes*, 3 Geo. L. Tech. Rev. 339, 339 (2019).
[74] Lindsey Wilkerson, Note, *Still Waters Run Deep(fakes): The Rising Concerns of "Deepfake" Technology and Its Influence on Democracy and the First Amendment*, 86 Mo. L. Rev. 407, 408 (2021).
[75] Princeton Legal Journal, *The High Stakes of Deepfakes*, https://legaljournal.princeton.edu/the-high-stakes-of-deepfakes-the-growing-necessity-of-federal-legislation-to-regulate-this-rapidly-evolving-technology/ (2023).
[76] Erik Gerstner, *Face/Off: "DeepFake" Face Swaps and Privacy Laws*, https://www.iadclaw.org/assets/1/17/Face_Off_-_DeepFake_Face_Swaps_and_Privacy_Laws.pdf?4179 (2020).

created by deepfakes may be caused by content that goes against societal values, violates religious taboos, or includes sexual content, such as revenge pornography.[77]

In addition to the complexities involving First Amendment and defamation laws in regulating deepfakes, state and federal governments have begun taking steps to mitigate potential harm. Currently, there is no federal legislation that addresses deepfake technology specifically. However, there have been recent developments in efforts to regulate deepfake technology. For example, in 2019, Congress passed the National Defense Authorization Act (NDAA), which requires the Director of National Intelligence to "report on the use of deepfakes by international governments, its ability to spread misinformation, and its potential impact on national security."[78] However, this Act primarily aims to protect the United States from deepfakes originating from outside the country and does not address the issues of deepfakes within the United States.[79]

Besides federal efforts, several states have enacted legislation addressing issues raised by deepfake technology. For example, Texas and California passed laws that ban the use of deepfake to influence upcoming elections.[80] The amendment of Texas's election code punishes a person who intends to "injure a candidate or influence the result of an election" by creating a deepfake video and publishing or distributing it within 30 days of an election.[81] California passed legislation that amended the state's election code to prohibit deepfakes published "within 60 days

---

[77] Marc Jonathan Blitz, *Article:Deepfakes And Other Non-Testimonial Falsehoods: When Is Belief Manipulation (Not) First Amendment Speech?*, 23 Yale J. L. & Tech. 160, 2020.
[78] S.1790 § 5709. *See also* Princeton Legal Journal, *The High Stakes of Deepfakes*, https://legaljournal.princeton.edu/the-high-stakes-of-deepfakes-the-growing-necessity-of-federal-legislation-to-regulate-this-rapidly-evolving-technology/ (2023).
[79] Princeton Legal Journal, *The High Stakes of Deepfakes*, https://legaljournal.princeton.edu/the-high-stakes-of-deepfakes-the-growing-necessity-of-federal-legislation-to-regulate-this-rapidly-evolving-technology/ (2023).
[80] Tex. Elec. Code Ann. § 255.004 (West 2019); Cal. Elec. Code § 20010 (West 2019). *See also* Lindsey Wilkerson, Note, *Still Waters Run Deep(fakes): The Rising Concerns of "Deepfake" Technology and Its Influence on Democracy and the First Amendment*, 86 Mo. L. Rev. 407, 408 (2021).
[81] Tex. Elec. Code Ann. § 255.004 (West 2019).

of an election" if they were distributed "with actual malice ... [and] with the intent to injure the candidate's reputation or to deceive a voter into voting for or against the candidate, unless the media includes a disclosure stating that the media has been manipulated."[82] California and Virginia passed laws to prohibit the creation and dissemination of nonconsensual deepfake pornography.[83] Maryland amended its statute prohibiting child pornography to include deepfake technology.[84]

Deepfakes and Text-to-image tools can be used to defame individuals in similar ways. For example, deepfake technology relies on a "base video" and pictures of the targeted individual to create deceptive videos.[85] Text-to-image tools, such as Generative Fill, can similarly be used to "swap" the person from the original image to the "base picture" and adjust the unnatural areas.[86] With the help of AI, both tools can defame someone by replacing the individual in a preexisting video or image, making it appear as though the targeted individual engaged in actions performed by someone else.

For defamation regulations, if the creator uses AI Text-to-image tools by face-swapping the victim to a "base image", similar to how deepfakes function, Text-to-image tools should be regulated as a subset of deepfakes. If the process for generating manipulated content is similar between deepfakes and Text-to-image tools, they will raise similar concerns to society. Even though regulations are still evolving to address the concerns related to deepfakes, it is likely that any similar concerns arising from Text-to-image tools will also be addressed by future deepfake regulations. State regulations already intervened to address issues of deepfakes in the context of

[82] Cal. Elec. Code § 20010 (West 2019).
[83] Va. Code Ann. § 18.2-386.2(A) (West 2019); Cal. Civ. Code § 1708.86 (West 2020).
[84] Md. Code Ann., Crim. Law § 11-208 (West 2019).
[85] Erik Gerstner, *Face/Off: "DeepFake" Face Swaps and Privacy Laws*, https://www.iadclaw.org/assets/1/17/Face_Off_-_DeepFake_Face_Swaps_and_Privacy_Laws.pdf?4179 (2020).
[86] PHLEARN, *How to Swap Faces in Photoshop Using AI Generative Fill*, YouTube (Sept. 29, 2023), https://www.youtube.com/watch?v=9YFDOoM7I5Q.

elections and pornography, and these regulations are anticipated to be adopted by other states in the foreseeable future. Therefore, it would be unnecessary to treat the Text-to-image face-swapping feature as a separate issue if it can be adequately regulated by future deepfake regulations.

B. Text-To-Image Features That Are Beyond Those Of Deepfakes

Generative Fill possesses additional features that could be used for defamation. These aspects of the Text-to-image tools should be subject to stricter regulations due to the lack of established regulations in this area. Beyond face-swapping the individual, Generative Fill can alter and manipulate the items on the image itself through text input to defame others. As illustrated in Part III, creators of Generative Fill can take an image of the targeted individual and select an area within the image, so that creators can change its background or a specific item based on text instructions. In this way, creators can defame the individual by portraying the person with an inappropriate object or situating the person in an unsuitable location. Alternatively, Generative Fill can take an upper-body image and expand it into a full-body image, potentially defaming the individual based on gestures or clothing. This type of image manipulation through text instructions involves more deliberate effort than simply face-swapping the person to the base image. To achieve the result, creators must consciously go through the editing steps by entering a text prompt that best describes their intended outcome. Due to the distinctions between deepfakes and Text-to-image tools, as much as deepfake regulations are evolving, the current deepfake regulations are inadequate to deter Text-to-image creators from creating defamatory images or to compensate the victims of defamation.

The deliberate image editing process using text instructions makes the current deepfake regulations inadequate. The victims deserve fewer obstacles in pursuing their defamation claims,

especially against the designers of the AI Text-to-image tools. As a potential solution, this paper proposes that the regulation should create a presumption that defamation's intent requirement is fulfilled when text input is employed in AI Text-to-image tools to defame others. For public figures, there should be a presumption of actual malice; for private individuals, there should be a presumption of negligence. The presumption should arise when the edited image depicts a specific individual falsely engaging in an offensive act, whether societally, sexually, religiously, or in other ways. This presumption of the fulfillment of the required intent could prevent defendants, especially the designers, from raising the intent requirement as a complete defense against defamation lawsuits. The presumption could be rebutted if the designer company can prove that its safety measures against defamation risks are adequate, and the burden of proof should be on the designer company. This new regulation will promote fairness because the designer company has better access to inside information on reasonable safety measures. It can also encourage the designer company to implement better safeguards in detecting and preventing misuse. The victims could, as a result of the presumption, have a better chance of prevailing and receiving the adequate compensation they deserved through actions against either the creator, designer, or both.

One could argue that the presumption creates a chilling effect on the designer companies. Technology companies might hesitate to incorporate Text-to-image tools in their products because of the potential defamation claims. While it is true that for public policy reasons, we want to encourage companies to invest in new technologies, the proposed regulation is designed to be narrow enough to not interfere with the development of new technologies. First, the regulation will cover content that is false and offensive. Also, the presumption does not automatically mean that the designer will be responsible for defamation liabilities when an image

involves elements generated through text prompts. The proposed regulation only aims to create a presumption that the intention requirement is fulfilled, and the plaintiff still needs to establish other elements of defamation.

One could also argue that creating images based on public figures creates entertainment values for society, and therefore, such a presumption might undermine the entertainment values of these images. The public might see these images as a way to "poke fun" at public figures.[87] This argument is not persuasive because the parody exception is still valid as a defense to parody. The standard of parody is whether or not the content can be reasonably understood as describing actual facts… or events.[88] If a reasonable person can discern that the image cannot be reasonably understood as depicting facts, it is meant to be a parody, and there will be no liability on the part of the creator or the designer. The presumption, in such cases, could at least help position the victims in a better position to receive compensation if it is clear that the work is not a parody.

VI. Conclusion

AI Text-to-image tools defy our traditional approach to image editing and our belief in visual content. Tools like Generative Fill have introduced uncertainty regarding the exact sources of an image. Text-to-image tools can be used to facilitate defamation, and our legal framework and regulations are not prepared to safeguard the victims. While deepfake regulations can offer some protection in cases involving Text-to-image tools, the capabilities of Text-to-image tools extend beyond those of deepfakes. To address these issues, regulations should create the presumption that manipulation of images using text inputs that create false and offensive

[87] Lindsey Wilkerson, Note, *Still Waters Run Deep(fakes): The Rising Concerns of "Deepfake" Technology and Its Influence on Democracy and the First Amendment*, 86 Mo. L. Rev. 407, 408 (2021).
[88] *Hustler Magazine v. Falwell*, 485 U.S. 46, 57 (1988).

material fulfills the required defamation intent. The goal is to deter creators from targeting individuals using Text-to-image technology and provide fair compensation to defamation victims.